NFDI4Chem, Chemistry Consortium in the NFDI

Deliverable D4.5.1

Report on FAIRness of datasets published by the community

TA4 – Metadata, Data Standards and Publication Standards

Lead Institution: Leibniz Institute of Plant Biochemistry (IPB Halle)

TA lead: Steffen Neumann (IPB Halle)
Subgroup lead(s): Annett Schröter (FSU Jena), Pascal Scherreiks (FSU Jena), David Rauh (IPB Halle)

Contributing partner(s): Pascal Scherreiks (FSU Jena), Johannes Liermann (JGU Mainz), Sonja Herres-Pawlis (RWTH Aachen), Tillmann Fischer (IPB Halle), Hans-Georg Weinig (GDCh, Frankfurt), Willis Muganda (GDCh, Frankfurt), Nicole Jung (KIT, Karlsruhe), Ann-Christin Andres (JGU Mainz), John Jolliffe (JGU Mainz), Theo Bender (JGU Mainz).

Main authors of this deliverable: David Rauh (IPB Halle), Tillmann Fischer (IPB Halle), Pascal Scherreiks (FSU Jena), Theo Bender (JGU Mainz), Steffen Neumann (IPB Halle).

Substantial contributions to this deliverable by: Nicole Parks (ITC-RWTH), John Jolliffe (JGU).

Funded by the German Research Foundation (DFG)
Project number: 441958208
Programm: Nationale Research Data Infrastructure

Content

Executive Summary

This deliverable is part of the activities in TA5's Measure 5.2 on awareness and Measure 4.5 of TA4 on FAIR assessment.

The aim of these activities were to

1. identify common antipatterns with opportunities for improvements,
2. provide examples of best practice of data description and publication, and
3. highlight these datasets to raise appreciation for good data publication.

In the *FAIR4Chem contest*, TA4 and TA5 collaborated to receive, evaluate and reward research datasets published by the wider chemistry community, which also allows for a qualitative assessment of the FAIRness of datasets published. Two entries to this competition in each of the years 2022 and 2023 were awarded the FAIR4Chem Award, which was handed over at the GDCh JCF Spring Symposium in each case.

For a quantitative assessment in a large-scale survey, we sampled the metadata for a large set of chemistry datasets and performed an automated FAIRness scoring with the F-UJI tool.

A lot of the FAIR principles depend on community standards for metadata that are not objectively computable such as F2 "Data are described with *rich* metadata.", I2 "(Meta)data use vocabularies that follow the FAIR principles", R1 "(Meta)data are *richly* described with a *plurality* of *accurate* and *relevant* attributes" or R1.3 "(Meta)data meet *domain-relevant* community standards". These aspects of the FAIR principles always require domain experts, as with the assessment of datasets for the FAIR4Chem Award, to assess the FAIRness.

Project objectives

This deliverable has contributed particularly to the **Objective 5.1** of TA5 to mediate a cultural change to the community and **Objective 4.3** of TA4 to increase the adoption of standards through interactions with the scholarly publication process.

Moreover, this deliverable has contributed to the following key objective in NFDI4Chem[1]:

Key Objective 4 to engage with the chemistry community in Germany to create awareness for, and foster the adoption of FAIR data management.

[1] C. Steinbeck, O. Koepler, F. Bach, S. Herres-Pawlis, N. Jung, J. C. Liermann, S. Neumann, M. Razum, C. Baldauf, F. Biedermann, T. W. Bocklitz, F. Boehm, F. Broda, P. Czodrowski, T. Engel, M. G. Hicks, S. M. Kast, C. Kettner, W. Koch, G. Lanza, A. Link, R. A. Mata, W. E. Nagel, A. Porzel, N. Schlörer, T. Schulze, H.-G. Weinig, W. Wenzel, L. A. Wessjohann, S. Wulle, NFDI4Chem - Towards a National Research Data Infrastructure for Chemistry in Germany, *RIO* **2020**, *6*, e55852, DOI: 10.3897/rio.6.e55852.

Detailed report on the deliverable

Background

The FAIRness of datasets is not a black-and-white matter, but rather a continuous range of how well a resource or dataset fulfils the individual criteria. Multiple approaches are available to assess datasets in this continuum of the FAIR[2] space.

Description of Work

The FAIR4Chem Contest and Award

The FAIR4Chem Award was established to celebrate and reward published chemistry research datasets that best meet the FAIR principles and thus make a significant contribution to increasing transparency in research and the reuse of scientific knowledge.

The awards were promoted via the NFDI4Chem web page, our mailing lists, newsletters (NFDI4Chem, GDCh, NFDI, NadCH), printed flyers, social media (Twitter, LinkedIn), on conferences, workshops and other in-person meetings as well as in email signatures of many people involved in NFDI4Chem (fig. 1).

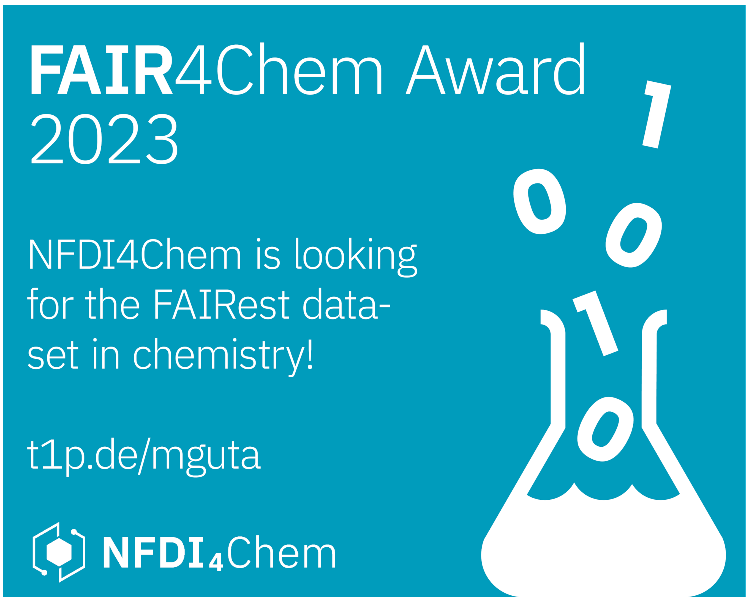

Figure 1: Banner used to advertise the FAIR4Chem Award in 2023.

Overview and winners in 2022 and 2023

The winners were selected in a two-stage process, using the criteria of the FAIR Data Self Assessment Tool[3] developed by the Australian Research Data Commons (ARDC) infrastructure, followed by a joint jury evaluation of the top-scoring submissions. The award

[2] M. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. *Sci Data* **2016**, *3*, 160018, DOI: 10.1038/sdata.2016.18.

[3] Australian Research Data Commons (ARDC), FAIR Data Self Assessment Tool https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/ .

includes a prize money of 500€, kindly provided by the Fonds der Chemischen Industrie (FCI).

The FAIR4Chem Award 2022 was open for submissions from 15th October 2021 to 15th December 2021. It honours the datasets of **Niels Krausch** and **Robert T. Giessmann** titled "*Collection of UV/Vis spectra acquired while monitoring reaction progress of thymidine phosphorolysis with varying reactant concentrations*", which was published in *Zenodo*[4], with its corresponding article published in *Processes*[5] and **Christopher Keßler** *et al.* for "*Supplementary material for 'Adsorption of Light Gases in Covalent Organic Frameworks: Comparison of Classical Density Functional Theory and Grand Canonical Monte Carlo Simulations*" published in DaRus[6], with its corresponding article published in *Microporous and Mesoporous Materials*[7].

The award and prize money were presented during the main program of the Spring Symposium of the youth organisation of the German Chemical Society (GDCh e.V.), the JungChemikerForum (JCF) on 25th March 2022 in Hannover, and published in the 6th NFDI4Chem Newsletter[8].

The second FAIR4Chem Award in 2023 was open for submissions from 1st September 2022 15th November 2022. The joint winners were **Johanna R. Bruckner** for her work titled "*Predictive design of ordered mesoporous silica with well-defined, ultra-large mesopores*", deposited in DaRUS[9], with its corresponding article published in Molecular Systems Design & Engineering[10]. The second winner was **Lena Daumann** for her dataset titled "*Modular Synthesis of New Pyrroloquinoline Quinone Derivatives*" stored in the Chemotion Repository [11], accompanied by a corresponding article published in *Synthesis*[12].

The awards were presented on 23rd March 2023 at the Spring Symposium of the GDCh JungChemikerForum (JCF) in Gießen. During the event, Prof. Daumann and Selina Itzigehl (on behalf of Prof. Bruckner) also delivered brief presentations on the research underlying their datasets. The dissemination of the award information took place on social media, the NFDI4Chem website[13], and the NFDI4Chem 10th Newsletter[14]. To raise further awareness of the FAIR4Chem Award and the importance of FAIR research data in chemistry, an interview with Lena Daumann was conducted and published on YouTube[15].

[4] N. Krausch, R. T. Giessmann, *Zenodo* **2019**, DOI: 10.5281/zenodo.3243351.

[5] R. T. Giessmann, N. Krausch, F. Kaspar, M. N. Cruz Bournazou, A. Wagner, P. Neubauer, M. Gimpel, *Processes* **2019**, *7*, 380, DOI: 10.3390/pr7060380.

[6] C. Kessler, J. Eller, J. Gross, N. Hansen, *DaRUS* **2021**, DOI: 10.1016/j.micromeso.2021.111263.

[7] C. Kessler, J. Eller, J. Gross, N. Hansen, *Microporous Mesoporous Mat.* **2021**, *324*, 111263, DOI: 10.1016/j.micromeso.2021.111263.

[8] Newsletter: 6th NFDI4Chem Newsletter.

[9] C. Vogler, S. Naumann, J. R. Bruckner, *DaRUS* **2022**, DOI: 10.18419/darus-2374.

[10] C. Vogler, S. Naumann, J. R. Bruckner, *Mol. Syst. Des. Eng.* **2022**, *7*, 1318, DOI: 10.1039/D2ME00107A.

[11] R. Janßen , V. A. Vetsova , D. Putz , P. Mayer , L. J. Daumann, *Chemotion* **2023**, DOI: 10.14272/collection/RAJ_2022-08-25.

[12] R. Janßen , V. A. Vetsova , D. Putz , P. Mayer , L. J. Daumann, *Synthesis* **2023**, *55*, 1000, DOI: 10.1055/s-0041-1738426.

[13] NFDI4Chem, FAIR4Chem Award 2023 laureates selected, FAIR4CHem Award 2023 the winners.

[14] Newsletter: 10th NFDI4Chem Newsletter.

[15] About FAIR data - NFDI4Chem www.youtube.com/watch?v=5ocj1wnI4E8 accessed on 2023-06-15.

The coverage of the FAIR4Chem Award and the announcement of the awardees garnered significant attention, with the news being among the most viewed posts by NFDI4Chem on social media. This highlights the importance and recognition of the Fair4Chem Award.

Details on the evaluation process

For both awards in 2022 and 2023, a google spreadsheet was developed to simplify the joint scoring of all submissions in an efficient manner, based on the ARDC tool. The results of this scoring led to the selection of candidates for the jury discussions.

For the FAIR4Chem Award 2022, we received 13 submissions, while three submissions did not meet the terms and conditions, as they were financed by NFDI4Chem, or linked to project web pages or online tools rather than datasets published in research data repositories. For the FAIR4Chem Award 2023, we received 6 submissions, while one did not meet the terms and conditions, as one of the co-authors was already awarded with the FAIR4Chem Award in 2022.

The average scores of all legitimate submissions with the ARDC tool in 2022 and also in 2023 gave an average score of 67%.

The highest score in 2022 was 87%, corresponding to the winner dataset of Christopher Keßler. The other winner, Krausch/Giessmann, had a score of 85%. This shows that awarded datasets had high scores with the ARDC tool. However, one other dataset also had a score of 87%, while examination of the dataset by the jury showed a lack in the description of the data, hence, the data were difficult to understand. Moreover, figures were added to the dataset, while the underlying data was not included.

Looking at all evaluated datasets, we observed clustering of scores around repositories. Indeed, with the ARDC assessment tool many points are awarded to the findability and accessibility aspect of the FAIR principles, such as “Does the dataset have any identifiers assigned?” or “Will the metadata record be available even if the data is no longer available?”. This in turn, as well as the accessibility and provision of APIs, highly relies on the repository chosen. As researchers have no direct control over the repository infrastructure, this choice does not reflect on the actual research data stored and how they were prepared before publishing.

Hence, we modified the ARDC assessment tool’s metrics prior to the evaluation of the submission for the FAIR4Chem Award 2023. We decreased the points for accessibility and halved the points for the provision of PIDs, as this is now common practice for most repositories. Furthermore, we added a measure of interoperability and questioned, whether metadata are part of the downloaded datasets (implementation of solutions such as BagIt or RO-crate) and increased the maximum number of points gained for the availability of analytical data in open formats, as the ARDC assessment tool issues only a few points on the latter aspect. Possibly, chemistry is a domain of research where the right data format is of particular importance, as there is a huge number of (mainly) proprietary but also some open formats available, especially for analytical data. Therefore, we also added a question for the aspect of reusability by asking for provenance information in analytical data files.

In 2023, the highest score was 79%, which was the winning dataset of Lena Daumann, while the other awarded dataset had a score of 70%. Again datasets with the highest ARDC score also proved to be selected in the final jury evaluation.

Large-scale evaluation with an automated FAIR assessment Workflow

For the large-Scale evaluation of datasets FAIR maturity[16] we used the automated FAIR data assessment tool F-UJI. The tool[17] allows to assess the FAIRness of research data objects given by a PID or URL using metrics of the FAIRsFAIR project[18].

Initially, we created a set of dataset DOIs to assess their metadata. The sample for analysis was obtained from DataCite[19], one of the main DOI registration agencies for research related digital objects that keep metadata records for these objects. Fig. 2 shows the full workflow used to produce the results of the FAIR assessment.

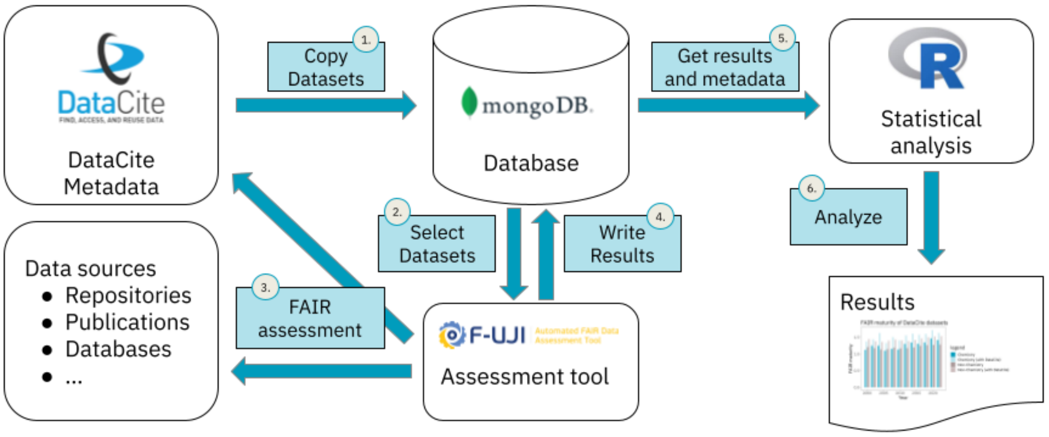

Figure 2: Automated FAIR assessment workflow using the F-UJI tool, and our additional infrastructure for large-scale analysis and data flow.

For analysis we gathered 31,395,796 public DataCite records (including records updated up to the 7th June 2022) by using the public DataCite REST API. A full text search[20] across all records was used to find chemistry related records. 11,891,880 of the records were describing research data sets and 664,343 of these were related to chemistry[21]. As input for the F-UJI tool only the DOI was needed from these records. To reduce the bias due to varying quality and quantity of metadata, the (mandatory) DOI prefix was used[22] to

[16] FAIR Data Maturity Model. Specification and Guidelines, *Zenodo* **2020**, DOI: 10.15497/rda00050.

[17] A. Devaraju, M. Mokrane, L. Cepinskas, R. Huber, P. Herterich, J. de Vries, V. Akerman, H. L'Hours, J. Davidson, M. Diepenbroek, *Data Sci. J.* **2021**, *20*, 1, DOI: 10.5334/dsj-2021-004.

[18] FAIRsFAIR, Fostering Fair Data Practices in Europe, URL: https://www.fairsfair.eu/.

[19] DataCite, URL: https://datacite.org/.

[20] Search terms: "chem", "chemistry", "chemical".

[21] Due to varying quantities of optional metadata in the DataCite records there may be some distortion towards higher FAIR scores in the chemistry data sets, because the chance of containing a keyword increases with the amount of additional metadata.

[22] We indeed identified a lack in the richness of metadata with DataCite's schema, where the domain of research can be added to the DOIs metadata within the tag *subject*, while this allows for free text values rather than providing a controlled list of possible domains and subjects and is often not filled.

differentiate the sources[23] of the data sets and to ensure data were also sampled from smaller sources. A sample of 18,366 of these records were selected for evaluation by the F-UJI tool. The sample was selected randomly with respect to a fair distribution over different DOI prefixes and years. This was necessary, because the amount of records with the same prefix can be in a range between only one and some thousands, and the number of new datasets increases every year. An unfiltered random sample would have given mainly new datasets from highly active uploaders. For all of the following statistics the average scores per DOI prefix were used to compensate for this effect.

There are two modes of assessment with the F-UJI tool: one is to directly use the landing page linked by the DOI to find the related data and metadata, and the other is to use information provided in the associated DataCite metadata. In the following analysis we compared both modes of FAIR evaluation (Fig. 3). These and the following results were also presented at the NFDI4Chem consortia meeting 3.0.

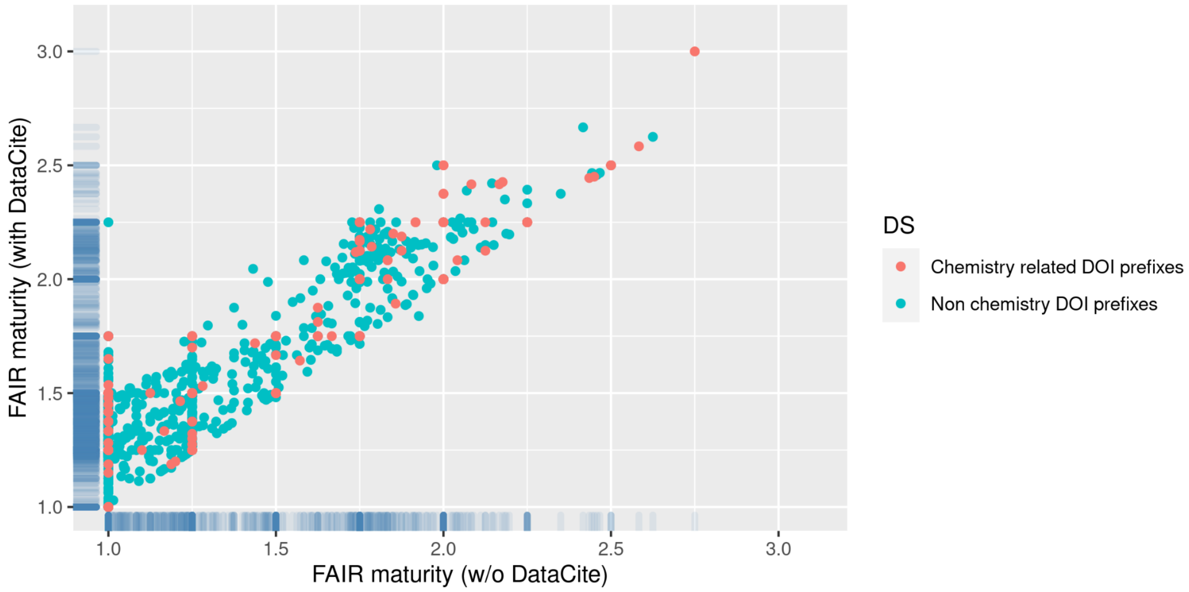

Figure 3: Comparison of the average maturity per DOI prefix for F-UJI runs with and without the inclusion of DataCite metadata for scoring FAIRness. Since the metadata were harvested from DataCite and therefore always have an identifier and a minimal set of metadata there are no datasets with a FAIR maturity level below 1.

FAIR maturity is a weighted score over all aspects of FAIRness and usage of the F-UJI tool will result in a maturity between 0 and 3[24,25]. The results show that including DataCite metadata in the analysis improves the maturity compared with the maturity assessed by

[23] DOI prefix only allows to deduce the registrant, which might provide repositories for several domains e.g. RADAR4Chem and RADAR4Culture. Indeed, DataCite Commons also uses the prefix to find all works published with a repository but do not use repository identifiers, e.g. provided by re3data.org. For academic publishing, journals with their scope and domain of interest use ISSN as identifiers within CrossRefs metadata schema, while the usage of repository identifiers is not common practice yet. Repository identifiers would ease the findability of chemistry related works also in DataCite metadata.
[24] A Devaraju, R. Huber, *Patterns* **2021**, *2*, 100370, DOI: 10.1016/j.patter.2021.100370.
[25] The maturity levels are defined in the source code of the F-UJI tool, URL: https://github.com/pangaea-data-publisher/fuji/blob/54c9b57fff5aa259b1ca3dd4fa693339a3e1404f/fuji_server/helper/metadata_mapper.py#L61

only using the landing page given by the DOI. The highest FAIR maturity levels of 3 were achieved by datasets published in PANGAEA.

If we look at the FAIR maturity with respect to the year of dataset creation (Fig. 4), we can see only a slight increase for newer datasets. The original hypothesis was a steady improvement over time, but as repositories continuously improve the FAIRness of their contents, even older datasets might have improved over time and may have had lower scoring at the time of their creation. Only some of the missing metadata in older datasets probably can not be recovered, so that they will always score lower. As explained earlier, the chemistry data sets tend to have a higher score than the non-chemistry datasets.

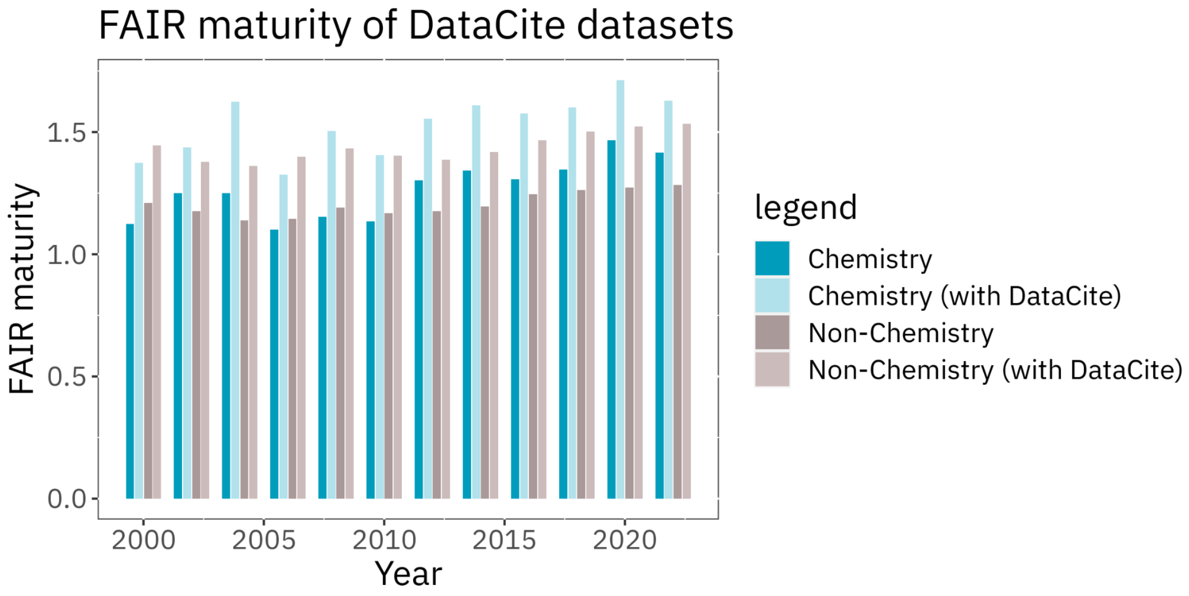

Figure 4: Average of FAIR maturity by year of dataset creation. See details in text for an explanation of the different record sets.

We had a closer look at the different aspects of FAIRness (Fig. 5). For findability we get the highest scores, which improve if we include DataCite metadata during analysis. Because the sample datasets were collected from DataCite in the first place, we can expect that they all have a DOI and can be searched for and found by their DOI, which will link it to the public landing page of the dataset. Links to other sites, included in the DataCite record might help to improve the result for the other metrics of the findability aspect. The scores for the other aspects – accessibility, interoperability and reusability – are lower and depend less on usage of the DataCite records.

By examining the results of the individual metrics used by the F-UJI assessment tool (table 1), we can find out more about the fields, where the publication of datasets has to be improved (Fig. 6). While overall scoring high in findability, some of the metadata records lack the identifier of the data they describe. For the other aspects most of the metrics score between 25% and 75%. Only the score for a recommended file format is very low. To draw any conclusion out of this, the chemistry datasets should be inspected in regards to the file format, because even non-recommended file formats may be accepted by the community as long as they meet other requirements for FAIRness, i.e. machine-readability or openness.

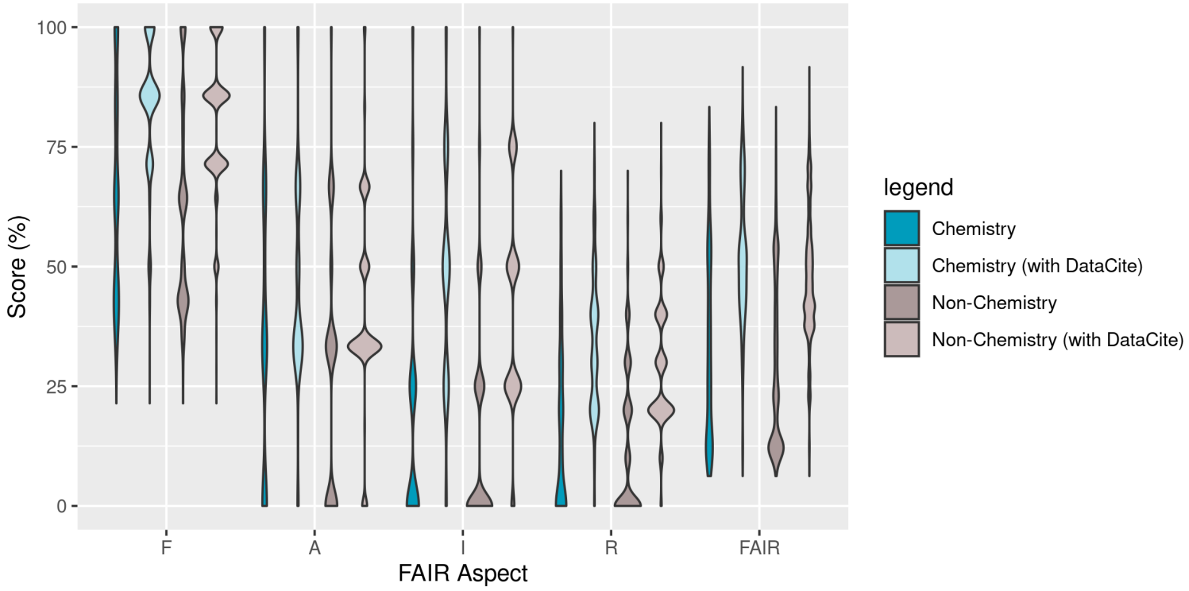

Figure 5: FAIRness of the datasets by aspect – which are findability, accessibility, interoperability and reusability. See details in text for an explanation of the different record sets.

In summary, the large scale assessment of datasets shows that many of the datasets get average scores in most of the metrics assessed by the F-UJI tool. Due to the sampling of datasets from DataCite – a provider for unique and persistent identifiers – scores in findability are high. Some of the FAIRsFAIR metrics were not collected by the F-UJI tool, others were only checked for existence of fields but not for their meaningfulness and content. Therefore the automated assessment can only give a rough overview about the FAIRness of datasets. Further studies could also include datasets which are related to publications but not accessible via DataCite to provide more insight in FAIRness of chemistry related datasets.

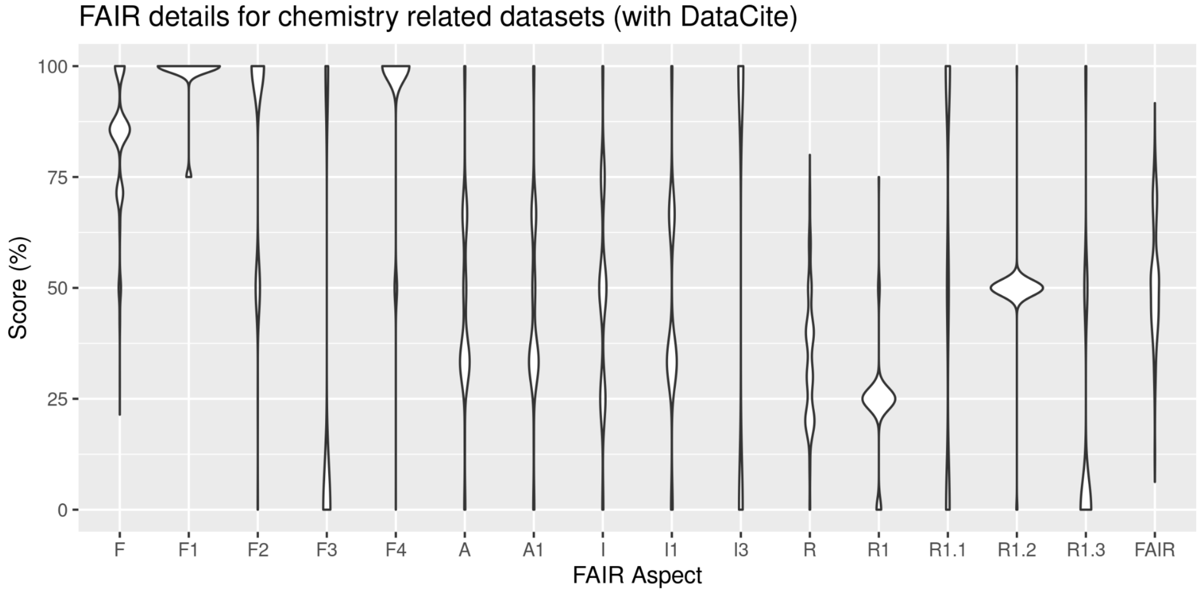

Figure 6: Scores for individual FAIRsFAIR metrics for the 18,366 sampled chemistry related datasets assessed by the F-UJI tool. See Table S1 for explanation of the individual FAIR aspects.

Evaluation of datasets from the FAIR4Chem Contest with F-UJI

In addition to the ARDC assessment tool's metrics, we also used the **F-UJI**[26] tool to evaluate the submission to the FAIR4Chem Contest. The average scores from this tool were 50% in both years, with the highest score in 2022 of 70% for two submissions not awarded. The awarded submissions in 2022 had F-UJI scores of 43% (Krausch/Griesmann) and 66% (Keßler). In 2023 the submission with the highest ARDC score and one of the winner datasets (Daumann) only had a F-UJI score of 45%, while the other winner (Bruckner) had a F-UJI score of 79%. This discrepancy between the automatically generated F-UJI scores and those from the ARDC questions reveals the limitations of complete automation in dataset FAIRness evaluation. Using an automated assessment tool can provide a starting point but does not yet substitute the manual evaluation process. A final discussion by a jury of domain experts from both research data management and a dataset discipline-specific context provides varying points of views and remains necessary.

Lessons learned

A lot of the FAIR principles depend on community standards for metadata that are not objectively computable such as F2 "Data are described with *rich* metadata.", I2 "(Meta)data use vocabularies that follow the FAIR principles", R1 "(Meta)data are *richly* described with a *plurality* of *accurate* and *relevant* attributes" or R1.3 "(Meta)data meet *domain-relevant* community standards".[27] These aspects of the FAIR principles always require domain experts to assess the FAIRness.

Next steps

We will initiate discussions on identifiers for research data repositories. Similarly, with the process now established, we will continue the FAIR4Chem awards as an additional incentive for the community, beyond Good Research Practices, to produce data that adheres to the FAIR principles and to show the benefits of these procedures in terms of effective research and personal reputation.

The Large-scale evaluation with the automated FAIR assessment tool F-UJI was a first test to load and process the metadata. The logical next step(s) are to improve the scientific usefulness of the metadata to capture discipline-specific information, and to improve the granularity such that metadata is available down to the data object or file level.

[26] Automated FAIR data assessment tool, URL: https://www.f-uji.net/.

[27] It's All in the Metadata: Marking Datasets FAIR, Mark A. Musen, CEDAR/Stanford University, RDA Thuringia, URL: https://www.youtube.com/watch?v=i1OliuOdQXY.

Appendices

Table 1: Brief listing of FAIR criteria assessed.

Metric	Description
F1	Data is assigned a globally **unique identifier**
F1	Data is assigned a **persistent identifier**
F2	**Metadata** includes **descriptive core elements** to support data findability.
F3	**Metadata** includes the **identifier** of the data it describes.
F4	**Metadata** is offered in such a way that it can be **retrieved by machines**.
A1	**Metadata** contains **access level and access conditions** of the data.
A1	**Metadata** is accessible through a **standardized** communication **protocol**
A1	**Data** is accessible through a **standardized** communication **protocol**
A2	**Metadata remains available**, even if the data is no longer available.
I1	**Metadata** is represented using a **formal knowledge representation language**.
I1	**Metadata** uses **semantic resources**.
I3	**Metadata** includes **links** between the **data** and its **related entities**.
R1	**Metadata** specifies the **content** of the data.
R1.1	**Metadata** includes **license information** under which data can be reused.
R1.2	**Metadata** includes **provenance** information about data creation or generation.
R1.3	**Metadata** follows a **standard** recommended by the target research community of the data.
R1.3	**Data** is available in a **file format** recommended by the target research community.